Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Árboles de regresión para el análisis de rating de avisos publicitarios del sector automotriz

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Milagros Doris PALOMINO MEZONES

ASESOR

Mg. Roberto Carlos FIESTAS FLORES

Lima, Perú

2021

Referencia bibliográfica

Palomino, M. (2021). *Árboles de regresión para el análisis de rating de avisos publicitarios del sector automotriz*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Milagros Doris Palomino Mezones
Tipo de documento de identidad	DNI
Número de documento de identidad	74278243
URL de ORCID	https://orcid.org/0000-0003-1250-6976
Datos de asesor	
Nombres y apellidos	Roberto Carlos Fiestas Flores
Tipo de documento de identidad	DNI
Número de documento de identidad	16744141
URL de ORCID	https://orcid.org/0000-0002-5582-0124
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Zoraida Judith Huamán Gutierrez
Tipo de documento	DNI
Número de documento de identidad	09890094
Miembro del jurado 1	
Nombres y apellidos	Carlos Alberto Jaimes Velasquez
Tipo de documento	DNI
Número de documento de identidad	42762905
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresa, Instituciones, Poblaciones locales, regionales y nacionales)

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Havas Media Perú S.A.C. País: Perú Departamento: Lima Provincia: Lima Distrito: San Isidro Avenida: Juan De Arona 151, Oficina. 703 Latitud: -12.096745 Longitud: -77.0321047
Año o rango de años en que se realizó la investigación	Julio 2021 – setiembre 2021
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL EN LA MODALIDAD VIRTUAL PARA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA

En Lima, siendo las 15:00 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional: Mg. Zoraida Judith Huamán Gutierrez (PRESIDENTA), Mg. Carlos Alberto Jaimes Velasquez (MIEMBRO) y el Mg. Roberto Carlos Fiestas Flores (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**ÁRBOLES DE REGRESIÓN PARA EL ANÁLISIS DE RATING DE AVISOS PUBLICITARIOS DEL SECTOR AUTOMOTRIZ**"**,** presentado por la señorita **Bachiller Milagros Doris Palomino Mezones***,* para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del trabajo de suficiencia, la Presidenta invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, la expositora mereció la aprobación de **SOBRESALIENTE,** con un calificativo promedio de **DIECISIETE (17)**

A continuación, los miembros del Jurado dan manifiesto que la participante **Bachiller Milagros Doris Palomino Mezones** en vista de haber aprobado la sustentación del Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 15:30 horas se levantó la sesión firmando para constancia la presente Acta.

Mg. Zoraida Judith Huamán Gutierrez
PRESIDENTA

Mg. Carlos Alberto Jaimes Velasquez
MIEMBRO

Mg. Roberto Carlos Fiestas Flores
MIEMBRO ASESOR

La Vicedecana de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, el titulando, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

RESUMEN

En la actualidad el mercado del sector automotriz está en crecimiento. La Asociación Automotriz del Perú, resalta el incremento en la venta de vehículos nuevos al cierre del primer semestre del 2021, pese a que este sector fue duramente golpeado con la pandemia hoy en día viene recuperándose y es una buena oportunidad para que la agencia de marketing optimice la compra de espacios publicitarios en el medio de televisión abierta, ya que, según un informe de la Asociación de Agencias de Marketing, Televisión es el medio con mayor inversión publicitaria.

Este trabajo de suficiencia profesional tiene como objetivo identificar un modelo estadístico para la toma de decisiones e identificar las variables más importantes a la hora de definir el rating. En la validación de datos se obtuvo que el coeficiente de determinación para la data de testeo fue de 0.77 y el RMSE 0.51. El mejor bloque para trasmitir los avisos publicitarios son el Estelar y Nocturno. En cuanto a variables más importantes encontramos la inversión, bloque horario, canal y genero de programa.

Palabras claves: Arboles de regresión, nodo, rating, avisos publicitarios.

ABSTRACT

Currently the automotive sector market is growing. The Automotive Association of Peru highlights the increase in the sale of new vehicles at the end of the first half of 2021, despite the fact that this sector was hard hit by the pandemic, today it is recovering and it is a good opportunity for the marketing agency optimize the purchase of advertising space in the broadcast television medium, since, according to a report by the Association of Marketing Agencies, Television is the medium with the highest advertising investment.

This work of professional sufficiency aims to identify a statistical model for decision-making and to identify the most important variables when defining the rating. In the data validation it was obtained that the determination coefficient for the test data was 0.77 and the RMSE 0.51. The best block to transmit the advertisements are the Star and Night. As for the most important variables, we find the investment, time block, channel and program genre.

Keywords: Regression trees, node, rating, advertisements.

ACTA DE SUSTENTACIÓN

Tabla de contenido

Índice de Figuras

Índice de Tablas

I. Introducción

La Asociación de Agencias de Medios, (2020) realizó un informe comparativo de dos periodos de la inversión publicitaria en el Perú, 2020 vs. 2019 (enero – diciembre), con el objetivo de brindar información relevante para la industria del marketing y comunicaciones. En este informe detallan que; la inversión publicitaria disminuyó 28% en el 2020, si bien el medio de Televisión abierta tuvo la mayor inversión, el medio Digital ocupó el segundo lugar muy cerca del primero y fue el único medio que creció (+4%). Vía Pública disminuyó su inversión publicitaria en -63%. Los medios Impresos tuvieron la mayor caída, Diarios fue el más afectado entre los medios masivos (-74%).

La Asociación Automotriz del Perú (AAP) (2021), resalta el incremento en la venta de vehículos nuevos al cierre del primer semestre del 2021. Según información de la Sunarp, la venta de Vehículos Livianos tuvo un marcado crecimiento de 154% respecto a junio del 2020 y en la comparativa del primer semestre del 2021 incrementó en 83% respecto al similar periodo del año pasado. Sostienen que el avance en ventas de este segmento de vehículos ha sido impulsado por la reactivación de la economía peruana, la cual viene mostrando dinamismo en la mayoría de sus actividades. Además, el informe de la APP menciona que durante los primeros seis meses del presente año los Vehículos Pesados comercializaron 138% adicional en comparación al similar periodo del 2020.

Uno de los objetivos de la agencia de medio es, mediante campañas de publicidad en los diferentes medios, lograr generar ventas a sus clientes. Vemos que el mercado del sector automotriz está incrementando en lo que va del año y mediante planificación en medios de comunicación, especialmente en Televisión Abierta ya que según la AAM es el medio con mayor inversión, se puede alcanzar este objetivo. Es primordial asegurar la correcta colocación de publicidad para los clientes, que su marca, producto, bien o servicio pueda ser visualizado por muchas personas y posicionarlo en el mercado. En TV, la audiencia es un indicador que nos mostrará cuantas personas fueron alcanzadas con la publicidad mostrada en este medio.

La importancia de este trabajo radica en poder predecir los niveles de audiencia, también conocidos como puntos de rating, en avisos publicitarios del sector automotriz, esto con la intención de conocer cuáles son los canales y/o programas que ofrecen mayor audiencia televisiva de tal forma llegar a más televidentes. Se aplicó el modelo de árboles de regresión para predecir la audiencia en el medio de Televisión Abierta.

Este trabajo se encuentra dividido en 6 capítulos, donde, el primer capítulo aborda la introducción, donde se explica de manera general el tema a desarrollar.
El segundo capítulo aborda información del lugar de trabajo donde se desarrolló la actividad.
En el tercer capítulo se describió a detalle la actividad realizada, mostrando como se organizó la información, finalidad, objetivos, problemática, metodología, procedimientos y resultados.
El cuarto capítulo abarca las conclusiones obtenidas.
El quinto capítulo muestra algunas recomendaciones basadas en las conclusiones obtenidas en el capítulo anterior.
El sexto capitulo incluye la bibliografía utilizada en este trabajo, los anexos y/o ilustraciones.

II. Información del lugar donde se desarrolló la actividad

- Institución donde se desarrolló la actividad

  La actividad se desarrolló en la agencia de medios Havas Media Group en el área de Measurement.

- Periodo de duración de la actividad

  La duración en el centro de labores donde se llevó a cabo la actividad fue desde el mes de noviembre 2020 a agosto 2021.

- Finalidad y objetivos de la entidad

  Finalidad

  Conocer el negocio de nuestros clientes potenciales y su entorno, a través de la investigación de mercados, de los desarrollos de decisión de compra de los consumidores y el reconocimiento de insights, generando así, soluciones de comunicación efectivas y visión estratégica.

  Objetivo

  Ser líderes en la construcción de marcas significativas y negocios exitosos. Sin duda nuestro éxito en la construcción de capacidades especializadas y el trabajo con Meaningful Brands son nuestros diferenciales para lograrlo.

- Razón social

  Havas Media Perú SAC

  RUC: 20417930079

- Dirección postal

  Av. Juan de Arona 151, San Isidro 15046

- Correo electrónico del profesional a cargo

  Correo: milagros.lib10@gmail.com

III. Descripción de la actividad

- **Organización de la actividad**

Para el desarrollo de este trabajo se realizaron diferentes actividades, las cuales detallaremos paso a paso:

- Entendimiento del negocio

  En esta fase inicial nos enfocamos en comprender la naturaleza del problema y su posible solución. Buscamos predecir los puntos de rating que se obtendrá en medios como Televisión Abierta.

- Entendimiento de los datos

  Esta fase comienza con la recolección de datos otorgado por la empresa, posterior a ello, revisamos las variables y vamos identificando algún conocimiento preliminar sobre los datos.

- Preparación de los datos

  En esta fase empezaremos empleando muestreo no probabilístico por conveniencia para obtener los datos. La tarea de preparación incluye la selección de los datos, limpieza, construcción de nuevas variables (en caso fuese necesario) y la integración de los mismo, de tal modo construir el conjunto de datos final.

- Modelado

  Durante esta fase se aplicó arboles de regresión al conjunto de datos obtenidos en la etapa anterior, el cual nos permitió predecir los puntos de rating obtenidos en el medio de televisión abierta.

- Evaluación

  En esta etapa final se evaluó si el modelo logra alcanzar la calidad suficiente, esto mediante el indicador de RMSE.

- **Finalidad y objetivos de la actividad**

Finalidad:

Elaborar un modelo estadístico para la toma de decisión al momento de elegir la compra de espacios publicitarios en televisión en avisos publicitarios del sector automotriz.

Objetivos Específicos:

Describir las variables que son más importantes a la hora de formar el rating.

Identificar el bloque horario que genere mayor rating.

Identificar el género del programa que genere mayor rating.

- **Problemática**

Palacios, (2020) en su trabajo de investigación que lleva por título "Análisis y predicción de las tendencias de venta en el mercado" busca aplicar una estrategia de predicción de mínima inversión basada en el uso de árboles de regresión. El autor realiza un análisis performance usando como indicador la métrica de precisión de los modelos implementados en datos prácticos enfocados a la comercialización y venta de bienes raíces. Una de los resultados rescatables del trabajo, muestra al modelo CART como el modelo con buenos resultados, logrando valores de precisión entre 80% y 94%. Este resultado proporciona al autor la certeza de un buen rendimiento de predicción y que el modelo aplicado es una forma factible para apoyar la administración de cualquier pequeña y mediana empresa.

Rodriguez, (2018) en su trabajo de investigación que lleva por título "Modelización y análisis de la calidad del aire en la ciudad de Oviedo" tiene como objetivo hallar un modelo adaptable a la resolución de problemas relacionados con la calidad del aire, empleando diversos modelos basados en el aprendizaje estadístico. Para lograr tal objetivo, el autor anota un set de datos experimentales como dióxido de azufre, monóxido de carbono, óxidos de nitrógeno, entre otros, desde el año 2013 hasta el 2015 en la ciudad de Oviedo. Luego de elaborar los modelos, el autor concluye que al comparar los valores de dióxido de nitrógeno (NO_2) observados y predichos aplicando la técnica de árbol de regresión obtiene un $R^2 = 0.75$, mientras que aplicando la técnica Red MLP $R^2 = 0.80$.

Maydana, (2021) en su trabajo de investigación que tiene por título "Elección del mejor para predecir el precio máximo de las acciones de Intel", busca comparar el modelo de árboles de regresión y regresión lineal múltiple. Empleó una muestra de 410 registros. La autora obtuvo como resultado que en el modelo de árbol de regresión obtiene un error cuadrático medio de 1.45 dólares, mientras que con la regresión lineal múltiple obtiene un

$R^2 = 0.97$ y un Error Estándar Residual de 0.2257 dólares. Al finalizar el trabajo de investigación, la autora llega a la conclusión que el mejor modelo para predecir el precio máximo de acciones es la regresión lineal múltiple con eliminación de datos atípicos.

Alcalá, (2017) en su trabajo de investigación llamado "Árboles de Regresión. Algunos algoritmos y extensiones a métodos de consenso" nos muestra el concepto de árboles de clasificación y métodos de agregación, tales como Bagging, Random Forest y Boosting. Después realiza la aplicación de un caso práctico empleando información de datos de jugadores de la liga española de los años 2015 y 2016, la muestra consta de 547 individuos y 38 variables. El objetivo principal es poner en práctica los métodos explicados. Como parte del modelado, el autor divide la data en training y test con 358 y 179 jugadores respectivamente. Para evaluar si el desempeño de los árboles es relativamente bueno lo compara con un modelo de regresión, el cual está formado con las mismas variables que los árboles. Llegando a la conclusión que el mejor método es Boosting con un $R^2 = 0.741$, superando al obtenido en el modelo de regresión (R^2: 0.68), mientras que solo el árbol de decisión obtuvo un R^2: 0.36.

- **Metodología, Procedimientos**

Tipo y diseño de la investigación

En este capítulo de la metodología lo primero que se debe tener en cuenta es el tipo de investigación que se realizó. Ya que establece los pasos a seguir del estudio, los métodos y técnicas que pueden ser empleados.

El enfoque es de tipo cuantitativo, ya que empleamos la técnica estadística, árboles de regresión, como un medio alcanzar los objetivos planteados, para modelar los datos y obtener resultados.

El diseño de esta investigación es no experimental, debido a que no se manipulan las variables ni los datos, solo se observan, tampoco hay intención de modificarlas.

La población de este trabajo fueron todos los avisos publicitarios emitidos en medios tradicionales como lo es la televisión abierta comprendido en el periodo de enero 2020 a julio 2021.

La muestra fue extraída por conveniencia, ya que se seleccionaron los avisos emitidos en el periodo de enero a julio del presente año, logrando obtener un conjunto de 2,646 registros.

Variables de investigación

Rating (Rat%) – Audiencia media

Es la división entre los televidentes medios de un evento y la población total del target en análisis, expresado en %. Asimismo, puede ser expresado como la división del Tiempo Medio de audiencia del evento y la duración total de éste. (Boletín Kantar Ibope Media).

$$Rating\% = \frac{Audiencia}{universo\ (target)} * 100$$

Marca

Es la denominación distintiva de un producto o servicio en un mercado determinado. La denominación que se registra corresponde al nombre comercial de la marca.

Consideraciones:

➢ Marcas Genéricas: Se utiliza para agrupar diferentes marcas que poseen características similares, cuya presencia en medios no representan un impacto considerable en su volumen e inversión.

➢ Multimarcas: Cuando un aviso incluye más de una marca se confirma con el cliente que ordena el aviso y se registra ambas marcas registrando primero la marca que ha sido ordenada por la agencia o central y en base a la primera marca registrada asignar el respectivo ítem/categoría/sector

Producto

Es el bien o servicio que ofrece una marca con el fin de satisfacer un deseo o una necesidad.

Atributos del producto que se deben evidenciar de manera explícita en el anuncio:

➢ Denominación: propia como un derivado de la marca

➢ Temporalidad: cuando el producto se trata de una promoción u oferta.

➢ Exclusividad: cuando el producto que se identifica en el aviso solo se obtiene en los canales de distribución de un único anunciante.

Duración

Duración Real del aviso en segundos (sólo para Tv y Cable).

Inversión

Valor monetario del aviso según lo publicado por el medio. De acuerdo con la configuración del sistema puede verse en soles o dólares. Así mismo, existen algunos tipos de avisos que muestran inversión estándar en función a criterios metodológicos.

Género.

Género del programa durante el cual se emitió el aviso (sólo para Tv, Cable y Radio).

Emisora

Nombre de la estación que emite la señal en el caso de tv, cable y radio y en el caso de Medios Impresos la denominación de los diarios, revistas y suplementos

Programa

Nombre del programa en el que se emitió el aviso (sólo para Tv, Cable y Radio).

Bloque Horario

Segmento horario donde se emitió el aviso reportado (sólo para Tv, Cable y Radio).

Clasificación y Regresión

Hay dos principales tipos de problemas de aprendizaje automático supervisados, denominados clasificación y regresión. En la clasificación, la finalidad es predecir una etiqueta de clase, que es una elección de una lista predeterminada de posibilidades. La clasificación a veces se divide en clasificación binaria, donde se busca distinguir entre exactamente dos clases, y clasificación multiclase la cual pretende distinguir entre más de dos clases. Para las tareas de regresión, la finalidad es predecir un número continuo o un número real. Muller C. & Guido (2017)

Árboles Clasificación y Regresión

Ahora que ya definimos la diferencia entre clasificación y regresión, nos enfocaremos en introducir el conceto de árboles. Los árboles de clasificación incluyen aquellos modelos en los que la variable dependiente es categórica y los árboles de regresión incluyen aquellos modelos en los que la variable predicha es continua. En adelante, nos orientaremos a introducir mayores conceptos sobre arboles de regresión, ya que esta técnica estadística nos ayudará a cumplir con los objetivos de este trabajo.

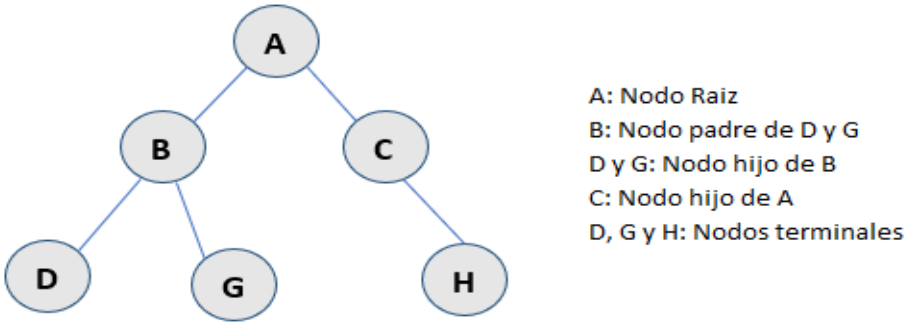
Árboles y conceptos generales

Los árboles de regresión son la subfamilia de árboles de decisión que se aplica cuando la variable respuesta es continua. En términos globales, en el train de un árbol de regresión, los datos se van dividiendo por nodos (separaciones) ocasionando la forma del árbol hasta lograr un nodo terminal. Cuando se busca pronosticar una nueva observación, se camina el árbol de acuerdo con el valor de sus predictores hasta llegar a uno de los nodos terminales. La predicción del árbol es la media de la variable respuesta de las observaciones de train que están en ese mismo nodo terminal.

Partes de un árbol

- Un nodo es la unidad en la cual se construye un árbol.
- Nodo interno (Root Node): denota una prueba sobre un atributo.
- Rama (Branch): corresponde un valor de atributo y representa el resultado de una prueba.
- Nodo terminal (Leaf Node): representa una etiqueta de clase o de distribución de clase.

Figura 1. Gráfico de un árbol



Teorema 1: La constante k que minimiza el valor esperado del error cuadrático es el valor medio.

Demostración:

$$\Phi(k) = E[(Y-k)^2]$$

Donde:

$$\begin{aligned}\Phi(k) = E[(Y-k)^2] &= \int_{-\infty}^{+\infty} (y-k)^2 f(y)dy \\ &= \int_{-\infty}^{+\infty} (y^2 - 2yk + k)^2 f(y)dy \\ &= \int_{-\infty}^{+\infty} y^2 f(y)dy - 2k\int_{-\infty}^{+\infty} y^2 f(y)dy + k^2\end{aligned}$$

Buscamos minimizar la función, por ello derivamos respecto a k buscamos igualar la derivada a 0:

$$0 - 2\int_{-\infty}^{+\infty} y^2 f(y)dy + 2k = 0$$

$$k = \int_{-\infty}^{+\infty} yf(y)dy = E[Y]$$

Por definición de $E[Y]$, supuesto finito.

Elaboración de Árboles de Regresión

Para la construcción del árbol, emplearemos las siguientes notaciones:

n: # de casos.

n_l: # de casos en la hoja l.

$\tilde{T}$: conjunto de hojas del árbol T.

T_t: subárbol del nodo t.

Y: Vector de la variable dependiente $Y = (y_1, y_2, \dots, y_n)^T$.

X: Matriz de las variables predictoras.

m_1: $\frac{1}{n_l}\sum_{i\in l} y_i$, media de la hoja l

$S(T)$: Suma del error cuadrático del árbol T definida como

$$S = \sum_{l\in\tilde{T}} \sum_{i\in l} (y_i - m_l)^2.$$

Algoritmo de construcción de un árbol

Referencia: X, Y, criterio de parada, criterio de nodo terminal.

Inicio: El nodo padre simboliza todo el espacio de variables independientes X. Se atribuye la media de los valores de Y, m_T. Calcular $S(T)$.

1er Paso: Calcular la suma del error cuadrático S(T) para todos los posibles cortes c de los nodos no terminales, si todos los nodos son terminales entonces FIN DEL ALGORITMO.

2do Paso: Elegimos el corte c que nos proporcione menor S.

3er Paso: Si obedece el criterio de parada y nodo terminal, aplicar el primer paso, en caso de no cumplirse, añadir 2 nodos no terminales y aplicar el primer paso.

FINAL DEL ALGORITMO

Como podemos preciar, el modelo toma una forma de búsqueda recursiva, pues realiza óptimas decisiones en cada paso ubicando el mejor corte c. Y es decreciente, ya que comienza en el nodo padre y sucesivamente divide el espacio de variables independientes.

El criterio de parada acostumbra a fijarse como un umbral δ que es la mejora mínima que se exige cada vez que se realiza un corte. En cuanto al principio de nodos terminales, es frecuente encontrarnos con condiciones como un mínimo número de datos en cada nodo, de tal forma evitar nodos con poca cantidad de datos.

1. Detallar criterios de parada y nodos terminales. A estos criterios se les acostumbra a llamar criterios de pre - poda.

2. Elaborar un árbol demasiado grande para después poder podarlo.

Importancia de las variables

En ejercicios de data mining, las variables independientes casi nunca son igual de importantes. Generalmente, solo unas limitadas variables tienen un dominio esencial en la variable endógena, pero la mayoría de las variables son insignificativas y pueden asimismo tener sentido no considerarlas dentro del modelo. Por eso, frecuentemente es de suma relevancia entender la significancia de cada variable de entrada para pronosticas nuestra variable respuesta. Los métodos de elaboración de Bagging y Random Forest nos admite construir un conjunto de normas para poder identificar la importancia de cada variable en el pronóstico final. Es por ello que se evalúa el error de la muestra. Luego, para cada variable de la muestra se permuta y se vuelve a calcular el error de la muestra permutada.

Resultados de la actividad

Entendimiento del problema

Las necesidades que en la actualidad aqueja a la empresa fueron explicadas en la introducción y planteamientos de objetivos.

Entendimiento de los datos

En este punto nos enfocamos en determinar las variables que nos resultaron más relevantes del conjunto de datos, esto con la intención de disminuir la dimensionalidad de nuestro dataset. Se realizó una muestra por conveniencia con información de enero a julio del presente año, logrando registrar 2,646 avisos publicitarios emitidos en televisión abierta a nivel nacional. Las variables que empleamos para el modelamiento de datos son; tipos de vehículos, marca de los vehículos, modelo, la duración en segundo del aviso publicitario emitido, el tipo de aviso publicitario, canal de emisión, inversión en dólares, género del programa, bloque horario, números de spots transmitidos y rating.

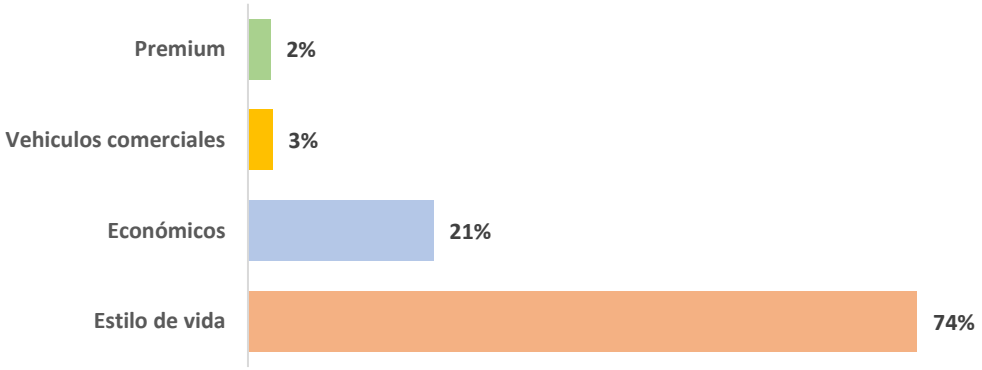
Preparación de los datos

Con nuestro subconjunto de variables y datos obtenidos en el paso anterior empezamos con el preprocesamiento de datos. Para este paso trabajaremos con el software Python versión 3.8.8. Debido a que los datos son obtenidos mediante un proveedor de la agencia, Kantar Ibope Media, este subconjunto no presenta datos perdidos.

Análisis Exploratorio de datos

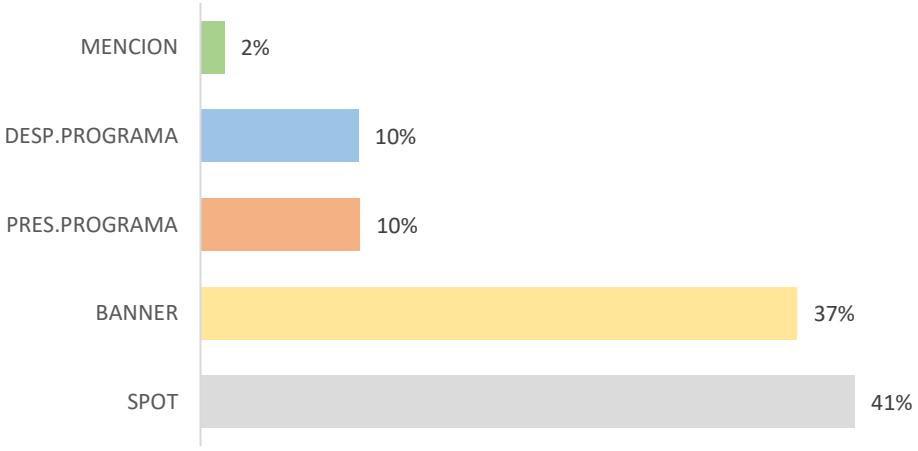
En este apartado, mostramos mediante gráficos y tablas el comportamiento univariado de nuestras variables.

Figura 2. Gráfico de barras según el tipo de vehículo



En el gráfico de barras podemos visualizar que la mayor cantidad de avisos publicitarios del sector automotriz emitidos en TV, pertenecen al tipo Estilo de Vida, estos son los modelos de autos usados para el día a día, comprende automóviles, camionetas y Pick Up. El segundo tipo con mayor frecuencia de emisiones son los Económicos, estos vehículos son los que en el mercado se les denomina "chinos", marcas como Baic, Changan, DFSK, Foton, Jac, entre otros, los encontramos en este tipo. Finalmente, los vehículos Comerciales y Premium son los que presentan menos emisión de avisos en TV.

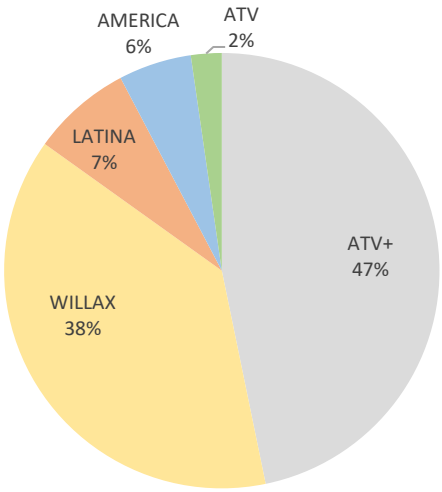
Figura 3. Gráfico de barras según el tipo aviso publicitario emitido



En el gráfico observamos que los Spots son los tipos de avisos con mayor número de apariciones en TV abierta (41%), seguido de Banner (37%). Las presentaciones y

despedidas de programas tienen 10% respectivamente y en menor porcentaje se ubican las Menciones de marca en vivo (2%).

Figura 4. Gráfico de sectores según canales de televisión abierta



En el gráfico de sectores visualizamos que los canales donde se emite con mayor frecuencia los avisos publicitarios del sector automotriz, con ATV+ con 47% y Willax 38%. Canales como Latina, América y ATV tienen 7, 6 y 2% respectivamente. La baja cantidad de avisos emitidos en estos 3 canales puede deberse al alto costo de los espacios publicitarios.

Tabla 1. Frecuencia de avisos publicitarios según el bloque horario

Bloque Horario	F	%
Femenino	1,146	54%
Estelar	526	25%
Matutino	299	14%
Nocturno	89	4%
Infantil	44	2%

De la tabla se observa que el bloque Femenino es quien lidera la emisión de avisos publicitarios con un 54% del share, seguido del bloque Estelar con 25%. Es coherente encontrar solo 2% de avisos trasmitidos en el bloque Infantil, ya que evidentemente este público no es objetivo dentro de las publicidades del sector automotriz.

Tabla 2. Genero de programas donde se emiten los avisos publicitarios del sector automotriz

Genero del Programa	F	%
Deportivos	1,408	67%
Noticieros	522	25%
Novelas-Magazine	77	4%
Concurso	55	3%
Otros	42	2%

De la tabla, el género Deportivo es quien lidera esta lista con un total de 67%, es muy común que los avisos publicitarios de vehículos aparezcan en programas deportivos ya que estos tienen una mayor audiencia de espectadores masculinos. Los programas con contenido de Noticias ocupan el segundo lugar con un 25% de emisiones de avisos. Finalmente, los programas con contenido de Novelas-Magazine, Concurso, Otros (Series, Comedias, Películas, Eventos, etc.) logran alcanzar el 9%.

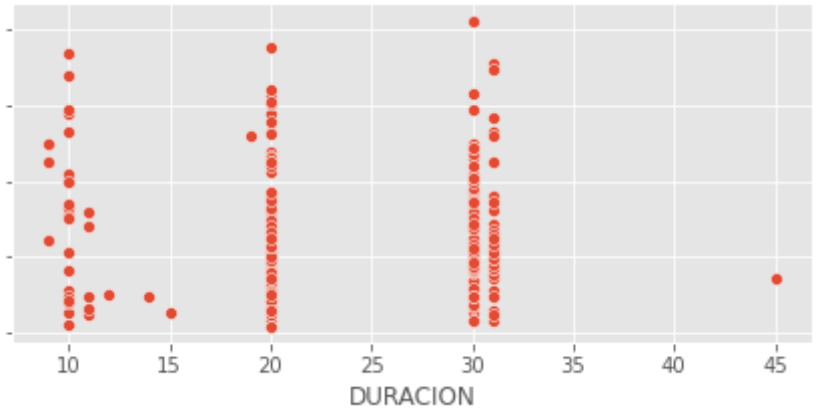
Tabla 3. Marca de vehículos con avisos publicitarios en TV

Marca de Vehículos	F	%
Toyota	519	25%
Kia	473	22%
Mg	395	19%
Mitsubishi	393	19%
Ford	171	8%

Lexus	52	2%
Volkswagen	31	1%
Dfsk	20	1%
Byd	18	1%
Renault	16	1%
Jac	14	1%
Seat	2	0%

De la tabla, las marcas que conforman el top 4 son Toyota, Kia, MG, Mitsubishi. Tener conocimiento de este top es fundamental ya que las marcas mencionadas son parte del set competitivo de nuestro cliente en la agencia. Por temas de practicidad procederemos a agrupar las marcas con menos de 10% en el share y lo renombraremos como “Otros”.

Figura 5. Gráfico de dispersión de la duración en segundos de los avisos emitidos en TV



En el grafico podemos visualizar que los avisos publicitarios que se emiten con mayor frecuencia son los de 20 y 30 segundos, estos por lo general suelen ser del tipo Spot. Por otro lado, tenemos avisos de entre 10 y 15 segundos, acá solemos tener presencia de los banners, menciones en vivo, presentación y despedida de programa. Además, podemos observar que solo se ha registrado un aviso con duración de 45 segundos en lo que va del periodo de enero a julio del 2021.

Modelamiento de los datos

Para efectos de trabajar de manera correcta en el software Python y con la librería sklearn, procedimos a convertir las variables de estudio en cualitativas nominal. La transformación de cada variable se puede visualizar en el ANEXO 1. También dividimos nuestra data en test (70%) y train (30%) para poder evaluar la capacidad predictiva del modelo.

Importancia de las variables

Con la intención de cumplir uno de nuestros objetivos específicos, deseamos conocer cuáles de las variables de nuestro conjunto de datos es más relevante a la hora de construir el árbol de regresión.

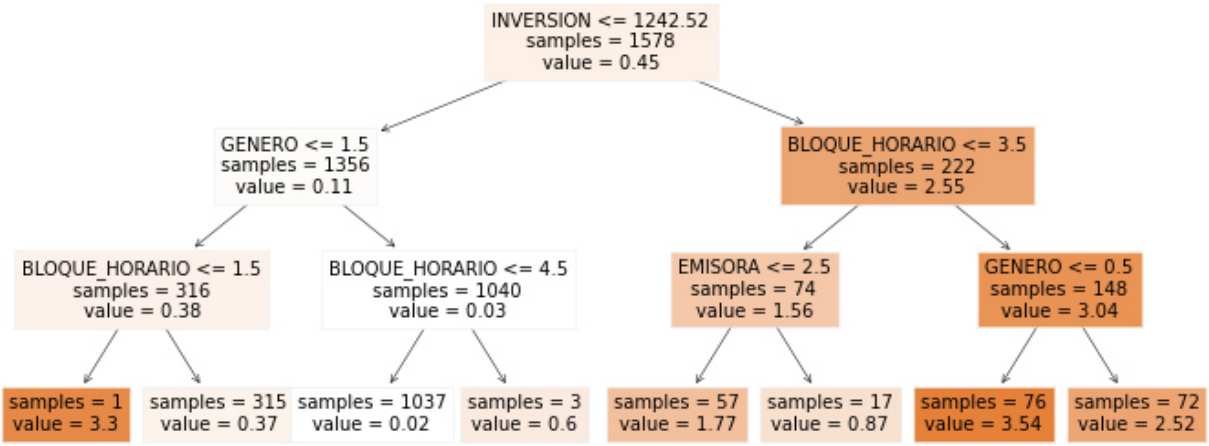
Tabla 4. Importancia de las variables predictoras

Variable	Importancia	
6	Inversión	0.852738
8	Bloque Horario	0.088302
7	Genero	0.051036
5	Emisora	0.007924
0	Tipo Vehículo	0
1	Marca	0
2	Producto	0
3	Duración	0
4	Tipo de Aviso	0
9	N° Spots	0

Las variables más significativas para la elaboración del modelo fueron Inversión, Bloque horario, genero del programa y Emisora (canal). La identificación de estas variables nos permitió realizar el árbol de regresión sin la necesidad de incluir las demás variables del conjunto de datos, de tal forma, poder conseguir mejores resultados al momento de evaluar el modelo propuesto.

Visualización del árbol de regresión

Figura 6. Árbol de regresión de avisos publicitarios emitidos en TV



Al momento de implementar el modelo, este ubicó la variable Inversión como mejor variable para empezar con la ramificación, vendría a ser nuestro nodo padre.

Realizamos la interpretación de la rama derecha; si la inversión publicitaria en TV es mayor a $. 1,242 y el Bloque horario es mayor a 3. 5, en otras palabras, si el bloque es Estelar o Nocturno y si el género del programa fuese deportivo, el aviso publicitario lograría obtener 3.54 puntos de rating, lo que equivale a alcanzar 35,400 hogares a nivel nacional

Por otro lado, si la inversión publicitaria en TV fuese menor a $. 1,242 y el género del programa fuese Deportico o Noticieros y el bloque horario Matutino, obtendremos 3.3 puntos de rating.

La forma de visualizar los resultados mediante el diagrama de árbol de decisión nos ayuda a comprender de manera práctica y sencilla.

Evaluación del modelo

Evaluaremos la capacidad predictiva de nuestro árbol, haciendo ejecución de nuestro código en Python, este código estará disponible en ANEXO II.

Para la evaluación de nuestro modelo emplearemos el RMSE, conocido por sus siglas en inglés como Root Mean Squared Error (Error Cuadrático Medio). Este Error fue calculado para la data de entrenamiento y testeo.

Tabla 5. Validación de resultados en train y test

	R^2	RMSE
Train	0.76	0.52
Test	0.77	0.51

El coeficiente de determinación (R^2) nos indica que el 76% del rating en televisión abierta es explicado por la inversión, bloque horario, genero del programa y canal donde se emite el aviso publicitario.

IV. Conclusiones

Este trabajo de suficiencia profesional ha descrito un acercamiento de los árboles de regresiones y como obtener un modelo apropiado en el contexto de predicción de puntos de rating obtenidos en televisión abierta en avisos publicitarios del sector automotriz. Se ha mostrado una metodología clara y generalizada para el procesamiento de datos, exploración y análisis de datos y elaboración de árboles de regresión.

El modelo elaborado nos muestra buenos resultados en cuanto a capacidad predictiva, dado que 76% del rating en televisión abierta es explicado por las variables de estudio.

Por otro lado, en cuestión a la formación del rating de avisos publicitarios en televisión abierta, hemos encontrado como las variables con mayor importancia para tener mayor alcance en la teleaudiencia, tienen que ver con la inversión publicitaria, el bloque horario y el canal donde se produce la emisión y el género del programa.

Además, si se busca optimizar la compra de espacios publicitarios y llegar a más televidentes, el mejor bloque horario es el Estelar, que va de 7 pm a 11:59 pm y el bloque Nocturno (12 pm a 2 am).

Finalmente, los tipos de programas que ayudaron a conseguir mayores puntos de rating en avisos publicitarios del sector automotriz fueron los deportivos seguido de noticieros. En el género deportivo se puede llegar a conseguir 3.3 hasta 3.54 puntos de rating.

V. Recomendaciones

Para clientes de la agencia de marketing enfocadas en el sector automotriz y donde se busque la compra de espacios publicitarios en televisión abierta, se recomienda realizarlo en los bloques horarios Estelar y Nocturno, de ese modo poder llegar a más televidentes.

Se recomienda al manager del área de Mesurament proponer a los supervisores el uso de modelos estadísticos para los clientes de la agencia, de tal forma encontrar insights enriquecedores en la data, ya que gran parte de la reportería esta basado en el enfoque descriptivo.

VI. Bibliografía

Asociación Automotriz del Perú. (2021). *Informe del Sector Automotriz a Junio 2021*.

Asociación de Agencias de Medios. (2020). *Barómetro de Medios 2020*.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (2017). Classification and Regression Trees. In *Chapter 8* (Issue January, pp. 1–7).

Ejea Carbonell, D. G. (2017). *Árboles de Regresión. Algunos algoritmos y extensiones a métodos de consenso*.

Maydana Huanca, A. R. (2021). *Elección del mejor modelo entre regresión lineal múltiple y árboles de regresión para predecir el precio máximo de las acciones de Intel en dunción al precio de apertura y volumen de ventas de acciones por día - 2019* (Issue 051).

Muller C., A., & Guido, S. (2017). *Introduction to Machine Learning with Python*. https://doi.org/10.1007/978-3-030-36826-5_10

Palacios Utreras, C. A. (2020). *Análisis y predicción de las tendencias de venta en el mercado usando árboles de regresión*.

Rodriguez Miranda, A. A. (2018). *Modelización y análisis de la calidad del aire en la ciudad de Oviedo (norte de España), mediante los enfoques PSO-SVM, red neuronal MLP y árbol de regresión*. 334. https://buleria.unileon.es/bitstream/handle/10612/7953/Tesis Alejandro Rodríguez Miranda.pdf?sequence=1&isAllowed=y

Anexo I: Transformación de variables

Tabla 6. Categorización de Tipo de vehículos

Nueva clasificación	Categoría
1	Estilo de vida
2	Premium
2	Vehículos comerciales
2	Económicos

Tabla 7. Categorización de Tipo de aviso publicitario

Nueva clasificación	Categoría
1	Banner
2	Spot
3	Mención
3	Pres. Programa
3	Desp. Programa

Tabla 8. Categorización según el canal

Nueva clasificación	Categoría
1	Latina
2	America
3	Atv
4	Atv+
5	Willax

Tabla 9. Categorización según el bloque horario

Nueva clasificación	Categoría
1	Matutino (06:00-11:59)
2	Femenino (12:00-15:59)
3	Infantil (16:00-18:59)
4	Estelar (19:00-23:59)
5	Nocturno (24:00-25:59)

Anexo II: Códigos en Python

```
import matplotlib.pyplot as plt
from sklearn import tree
from sklearn.metrics import mean_squared_error, r2_score
datos = pd.read_excel('D:/Titulación/EJEMPLO3.xlsx')
datos.head(5)
datos.shape
#Visualizaremos los datos de cada variable
sb.factorplot('TIPO_VEHICULO',data=datos,kind="count")
sb.factorplot('CATEGORIA',data=datos,kind="count")
sb.factorplot('TIPO_AVISO',data=datos,kind="count")
sb.factorplot('EMISORA',data=datos,kind="count")
sb.factorplot('BLOQUE',data=datos,kind="count")
sb.factorplot('MARCA',data=datos,kind="count" , aspect=3)
datos_autos = pd.read_excel('D:/Titulación/EJEMPLO3_1.xlsx')
datos_autos.head(5)
# Dividimos los datos en train y test
X_train, X_test, y_train, y_test = train_test_split(
                                   datos_autos.drop(columns = "RATING"),
                                   datos_autos['RATING'],
                                   random_state = 123
                                )
# Creamos el modelo
modelo = DecisionTreeRegressor(
           max_depth         = 3,
           random_state      = 123
         )

# Entrenamiento del modelo
modelo.fit(X_train, y_train)
```

```
# Estructura del árbol creado
fig, ax = plt.subplots(figsize=(12, 5))
print(f"Profundidad del árbol: {modelo.get_depth()}")
print(f"Número de nodos terminales: {modelo.get_n_leaves()}")

plot = plot_tree(
          decision_tree = modelo,
          feature_names = datos_autos.drop(columns = "RATING").columns,
          class_names   = 'RATING',
          filled        = True,
          impurity      = False,
          fontsize      = 10,
          precision     = 2,
          ax            = ax
      )
##importancia de predictores
importancia_predictores = pd.DataFrame(
                        {'predictor': datos_autos.drop(columns = "RATING").columns,
                         'importancia': modelo.feature_importances_}
                        )
print("Importancia de los predictores en el modelo")
importancia_predictores.sort_values('importancia', ascending=False)

# Error de test del modelo inicial
predicciones = modelo.predict(X = X_test)
rmse = mean_squared_error(
       y_true  = y_test,
       y_pred  = predicciones,
       squared = False
      )
print(f"El error (rmse) de test es: {rmse}")
```

```
# Error de test del modelo inicial
predicciones = modelo.predict(X = X_train)
rmse = mean_squared_error(
        y_true  = y_train,
        y_pred  = predicciones,
        squared = False
       )
print(f"El error (rmse) de train es: {rmse}")
```